# Project1 ANES_vote Shuqi Yu

Shuqi Yu

1/26/2021

## 1. Introduction

The *American National Election Studies* (ANES) are surveys of voters in the U.S. on the national scale. For each presidential election since 1948, ANES collects responses from respondents both before and after the election. The goal of ANES is to understand political behaviors using systematic surveys. ANES's data and results have been routinely used by news outlets, election campaigns and political researchers.

The *Time Series Cumulative Data* of ANES include answers, from respondents from different years, on selected questions that have been asked in three or more ANES' *Time Series* studies. Tremendous amount of efforts have been put into data consolidation as variables are often named differently in different years.

A rule of thumb for analyzing any data set is to understand its study design and data collection process first. You are strongly encouraged to read the *codebooks*.

## 2. Access to ANES Data

### Step 2.1: Register to access ANES dataset.

To access the data, you should register at ANES's website and accept its terms of use, especially committing to "use these datasets solely for research or statistical purposes and not for investigation of specific survey respondents."

### Step 2.2: Download the ANES Time Series Cumulative Data

Once you are logged into ANES's website, you should be able to download the data file. You can use ASCII, DTA or SAV. In this notebook, we use the *SAV* format. The downloaded file is a zip file, you should move all unzipped files into the `data` folder of your project 1's local folder.

## 3. Data processing for this R Notebook.

The following code blocks prepare a processed data set and save it in the `output` folder. The data folder should be only used for storing raw `data`. All processed data should be saved in the `output` folder. The notion here is that one can delete files from the output folder and reproduce them by re-running the codes.

#### Step 3.1 Checking `R` packages for data processing

From the packages' descriptions:

- `tidyverse` is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures;
- `haven` enables `R` to read and write various data formats used by other statistical packages. `haven` is part of the `tidyverse`.
- `devtools` provides a collection of package development tools.
- `RColorBrewer` provides ready-to-use color palettes.
- `DT` provides an R interface to the JavaScript library DataTables;
- `ggplot2` a collection of functions for creating graphics, based on The Grammar of Graphics.

### Step 3.2 Import raw ANES data

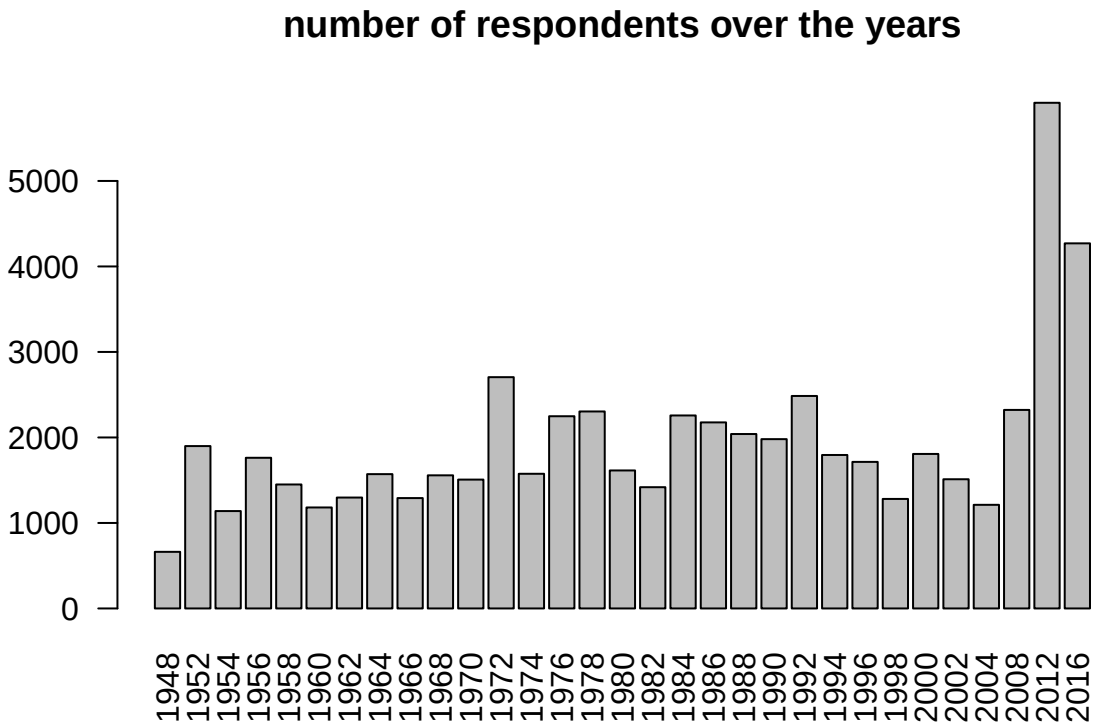
We will be working with the SAV format of the raw ANES data, downloaded from this page, once you are registered *and* logged in. This is a saved data file from SPSS. We will use the `read_sav` function from the `haven` package.

Read more about importing SPSS data into R.

```
library(haven)
anes_dat <-
    read_sav("../data/anes_timeseries_cdf.sav")
```

Some basic data summaries: there are 59944 respondents and 1029 variables.

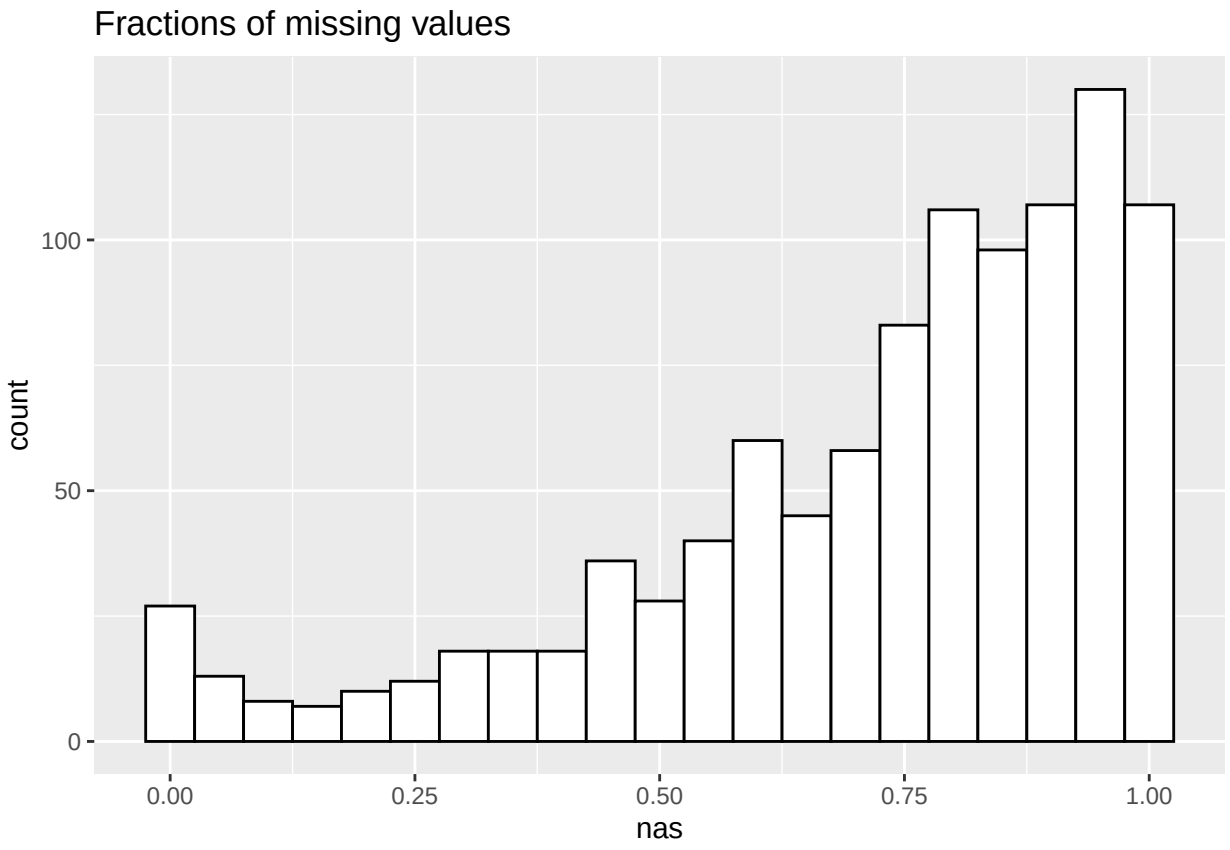
```
barplot(table(anes_dat$VCF0004),
        las=2,
        main="number of respondents over the years")
```




Some variables are asked nearly all the years and some are asked only a few years.

```
anes_NAs=anes_dat%>%
  summarise_all(list(na.mean=function(x){
                              mean(is.na(x))}))
anes_NAs=data.frame(nas=unlist(t(anes_NAs)))
```

```
ggplot(anes_NAs, aes(x=nas)) +
  geom_histogram(color="black",
                 fill="white",
                 binwidth=0.05)+
  labs(title="Fractions of missing values")
```




**Step 3.3 Process variables for analysis**

In the following, we will create a few variables for this tutorial. These variables were selected based on their description in the ANES codebook. You are encouraged to look them up and read about how they were prepared.

First let's look at our data. One advantage of using the SPSS SAV data is that the values are *labelled*. By converting labelled data to factor, you can easily reveal the responses encoded by numbers. We selected four variables for subsequent analysis and save the filtered data sets to the `output` folder.

```
Election_years=as.character(seq(1952, 2016, 4))

anes_use=anes_dat%>%
  mutate(
    year=as_factor(VCF0004),
    turnout=as_factor(VCF0703),
    vote=as_factor(VCF0706),
    race=as_factor(VCF0105a),
    gender=as_factor(VCF0104)
  )%>%
```

```
  filter(year %in% Election_years)

library(data.table)

data.table(anes_use%>%
             select(year, turnout, vote, race, gender)%>%
             filter(!is.na(turnout))%>%
             sample_n(30))
```

```
##     year                                turnout
##  1: 1992                3. Voted (registered)
##  2: 1976                3. Voted (registered)
##  3: 1960                3. Voted (registered)
##  4: 1976                3. Voted (registered)
##  5: 1968                3. Voted (registered)
##  6: 2016                3. Voted (registered)
##  7: 2012                3. Voted (registered)
##  8: 1952                3. Voted (registered)
##  9: 2016                3. Voted (registered)
## 10: 1980                3. Voted (registered)
## 11: 2016 1. Not registered, and did not vote
## 12: 2012 1. Not registered, and did not vote
## 13: 1952 1. Not registered, and did not vote
## 14: 1952                3. Voted (registered)
## 15: 2000     2. Registered, but did not vote
## 16: 1972                3. Voted (registered)
## 17: 2012                3. Voted (registered)
## 18: 1992 1. Not registered, and did not vote
## 19: 2012                3. Voted (registered)
## 20: 1992                3. Voted (registered)
## 21: 1984                3. Voted (registered)
## 22: 1984     2. Registered, but did not vote
## 23: 1968     2. Registered, but did not vote
## 24: 2012     2. Registered, but did not vote
## 25: 1964                3. Voted (registered)
## 26: 1952     2. Registered, but did not vote
## 27: 1960                3. Voted (registered)
## 28: 1964                3. Voted (registered)
## 29: 2016                3. Voted (registered)
## 30: 2008                3. Voted (registered)
##     year                                turnout
##                                                                vote
##  1:                                                   1. Democrat
##  2:                                                 2. Republican
##  3:                                                   1. Democrat
##  4:                                                 2. Republican
##  5:                                                 2. Republican
##  6:                                                 2. Republican
##  7:                                                 2. Republican
##  8:                                                   1. Democrat
##  9:                                                 2. Republican
## 10:                                                 2. Republican
## 11:                                                         <NA>
## 12: 7. Did not vote or voted but not for president (exc.1972)
```

```
## 13: 7. Did not vote or voted but not for president (exc.1972)
## 14:                                                1. Democrat
## 15: 7. Did not vote or voted but not for president (exc.1972)
## 16:                                              2. Republican
## 17: 7. Did not vote or voted but not for president (exc.1972)
## 18: 7. Did not vote or voted but not for president (exc.1972)
## 19:                                                1. Democrat
## 20:     3. Major third party candidate (Wallace 1968/Anderson
## 21:                                                       <NA>
## 22: 7. Did not vote or voted but not for president (exc.1972)
## 23: 7. Did not vote or voted but not for president (exc.1972)
## 24: 7. Did not vote or voted but not for president (exc.1972)
## 25:                                                1. Democrat
## 26: 7. Did not vote or voted but not for president (exc.1972)
## 27:                                              2. Republican
## 28:                                                1. Democrat
## 29:                                                1. Democrat
## 30:                                              2. Republican
##                                                          vote
##                                                          race    gender
##  1:                              2. Black non-Hispanic (1948-2012) 2. Female
##  2:                              1. White non-Hispanic (1948-2012) 2. Female
##  3:                              1. White non-Hispanic (1948-2012) 2. Female
##  4:                              1. White non-Hispanic (1948-2012) 2. Female
##  5:                              1. White non-Hispanic (1948-2012) 2. Female
##  6:                              1. White non-Hispanic (1948-2012) 2. Female
##  7: 4. American Indian or Alaska Native non-Hispanic (1966-2012)   1. Male
##  8:                              1. White non-Hispanic (1948-2012) 2. Female
##  9:                              1. White non-Hispanic (1948-2012)   1. Male
## 10:                              1. White non-Hispanic (1948-2012) 2. Female
## 11:                              1. White non-Hispanic (1948-2012)   1. Male
## 12:         6. Other or multiple races, non-Hispanic (1968-2012)   1. Male
## 13:                                                          <NA>      <NA>
## 14:                              1. White non-Hispanic (1948-2012) 2. Female
## 15:                                       5. Hispanic (1966-2012) 2. Female
## 16:                              1. White non-Hispanic (1948-2012)   1. Male
## 17:         6. Other or multiple races, non-Hispanic (1968-2012) 2. Female
## 18:                              1. White non-Hispanic (1948-2012)   1. Male
## 19:                                       5. Hispanic (1966-2012)   1. Male
## 20:                              1. White non-Hispanic (1948-2012)   1. Male
## 21:                              1. White non-Hispanic (1948-2012) 2. Female
## 22:                              1. White non-Hispanic (1948-2012) 2. Female
## 23:                              1. White non-Hispanic (1948-2012) 2. Female
## 24:                              2. Black non-Hispanic (1948-2012)   1. Male
## 25:                              1. White non-Hispanic (1948-2012)   1. Male
## 26:                              1. White non-Hispanic (1948-2012)   1. Male
## 27:                              1. White non-Hispanic (1948-2012) 2. Female
## 28:                              1. White non-Hispanic (1948-2012) 2. Female
## 29:                              1. White non-Hispanic (1948-2012) 2. Female
## 30:                              1. White non-Hispanic (1948-2012)   1. Male
##                                                          race    gender
```

```
anes_use = anes_use%>%select(year, turnout, vote, race, gender)
```

```
save(anes_use, file="../output/data_use.RData")
```

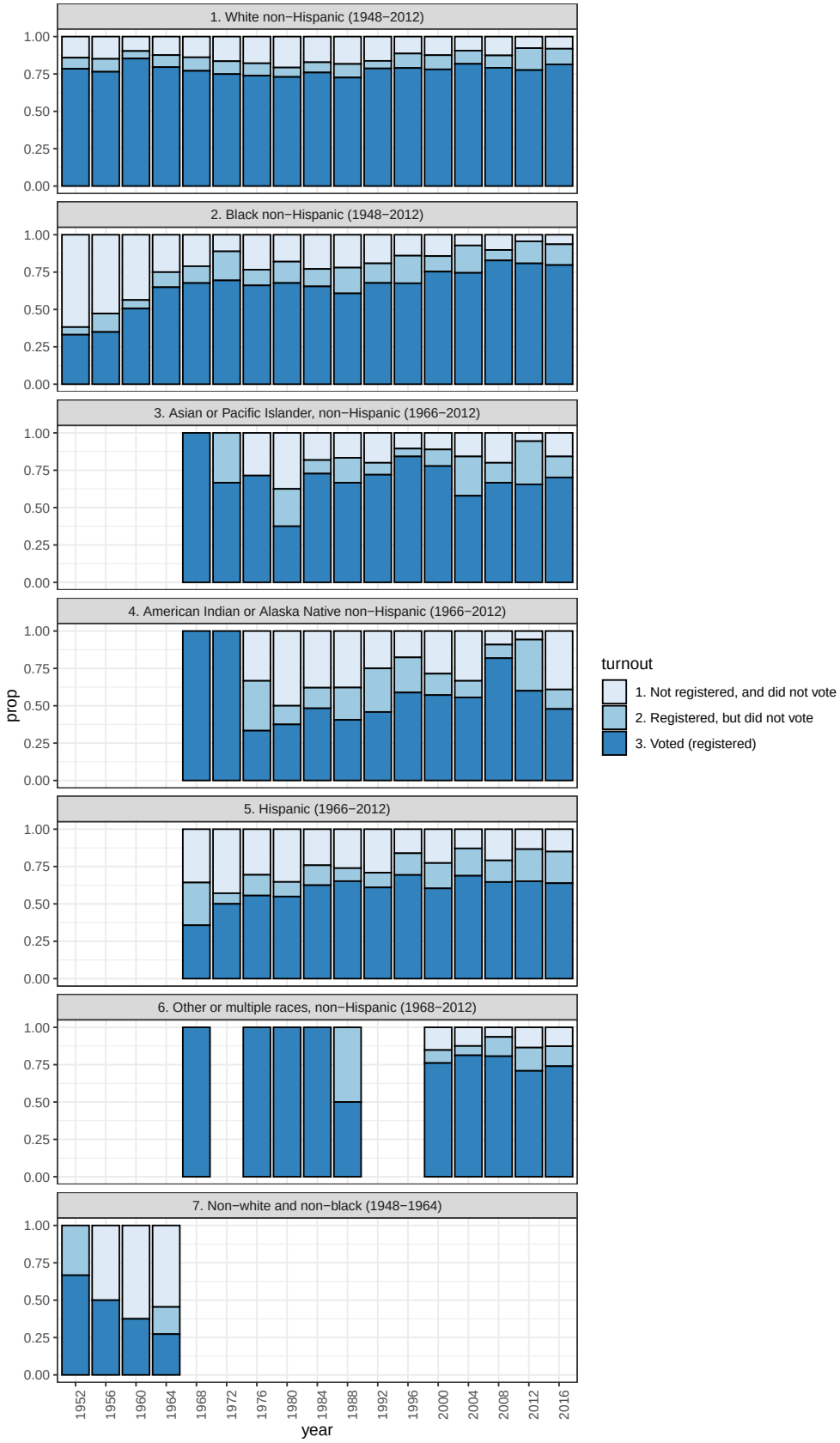
# 4. A simple analysis

## 4.1 Who went to vote in the election?

First, we would like to see whether different racial groups have different turnout rates on the election day.

```
load(file="../output/data_use.RData")
anes_to_race_year = anes_use %>%
  filter(!is.na(race) & !is.na(turnout))%>%
  group_by(year, race)%>%
  count(turnout)%>%
  group_by(year, race)%>%
  mutate(
    prop=n/sum(n)
  )

ggplot(anes_to_race_year,
       aes(x=year, y=prop, fill=turnout)) +
  geom_bar(stat="identity", colour="black") + facet_wrap(~race, ncol=1) + theme_bw()+
  theme(axis.text.x = element_text(angle = 90))+
  scale_fill_brewer(palette="Blues")+
  labs(title="How did different racial groups participated in the election \n over the years?")
```

How did different racial groups participated in the election over the years?

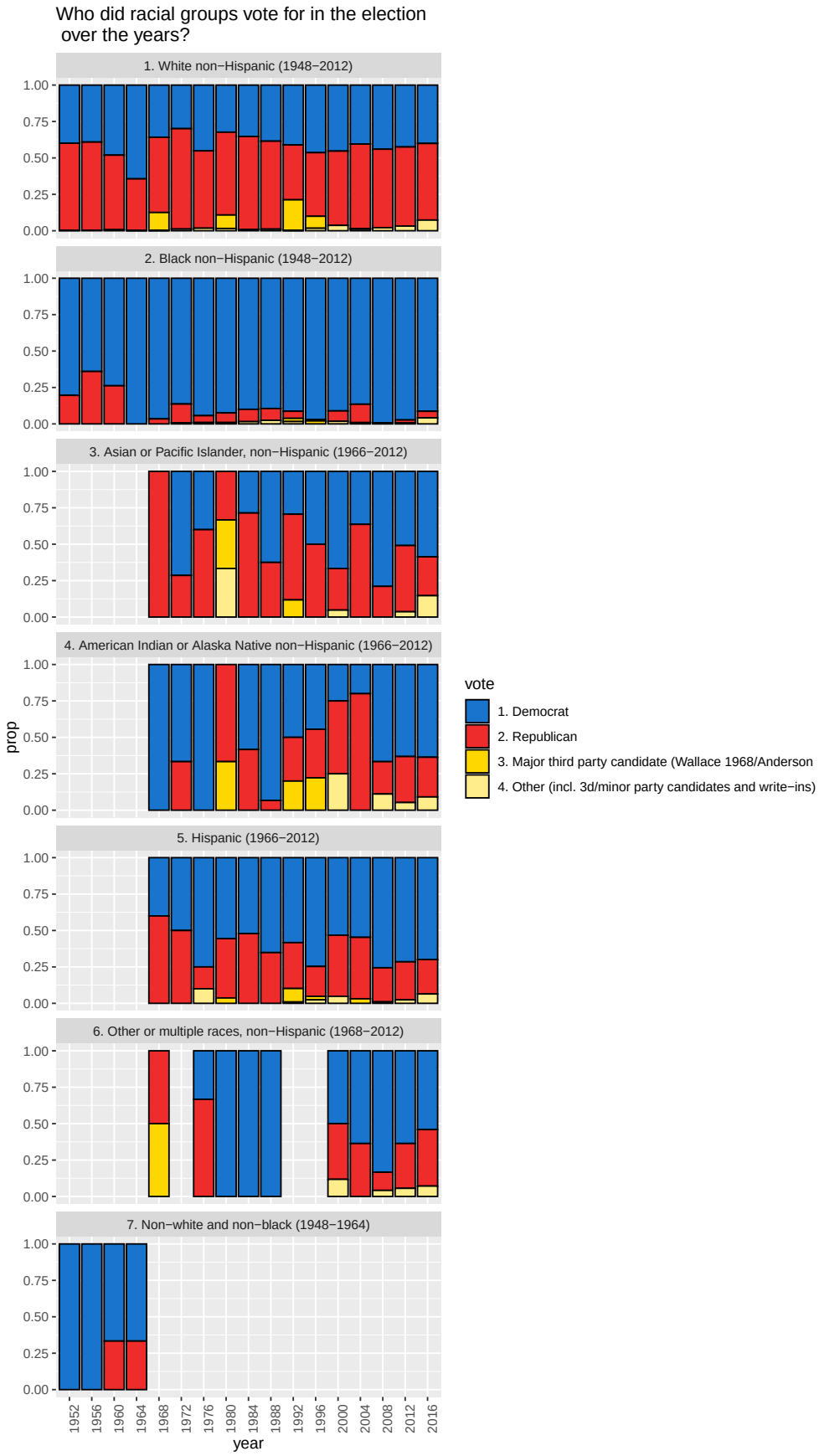


Wait a minute, this is not what we saw in the news!

Looks like other people have noticed it too and wrote a whole paper about this.

### 4.2 Who did they vote for in the election?

```
anes_vote_race_year = anes_use %>%
  filter(!is.na(race) & !is.na(vote))%>%
  filter(vote!="7. Did not vote or voted but not for president (exc.1972)")%>%
  group_by(year, race)%>%
  count(vote)%>%
  group_by(year, race)%>%
  mutate(
    prop=n/sum(n)
  )
#%>%
#  filter(vote == "1. Democrat" | vote == "2. Republican")

ggplot(anes_vote_race_year,
       aes(x=year, y=prop, fill=vote)) +
  geom_bar(stat="identity", colour="black")+
  scale_fill_manual(values=c("dodgerblue3", "firebrick2", "gold1", "lightgoldenrod1"))+
  facet_wrap(~race, ncol=1) +
  theme(axis.text.x = element_text(angle = 90))+
  labs(title="Who did racial groups vote for in the election \n over the years?")
```


Who did racial groups vote for in the election over the years?
1. White non−Hispanic (1948−2012)
2. Black non−Hispanic (1948−2012)
3. Asian or Pacific Islander, non−Hispanic (1966−2012)
4. American Indian or Alaska Native non−Hispanic (1966−2012)
5. Hispanic (1966−2012)
6. Other or multiple races, non−Hispanic (1968−2012)
7. Non−white and non−black (1948−1964)
prop
1.00
0.75
0.50
0.25
0.00
1952
1956
1960
1964
1968
1972
1976
1980
1984
1988
1992
1996
2000
2004
2008
2012
2016
year
vote
1. Democrat
2. Republican
3. Major third party candidate (Wallace 1968/Anderson
4. Other (incl. 3d/minor party candidates and write−ins)

## 5. The voting relationship with thier family burden and duty

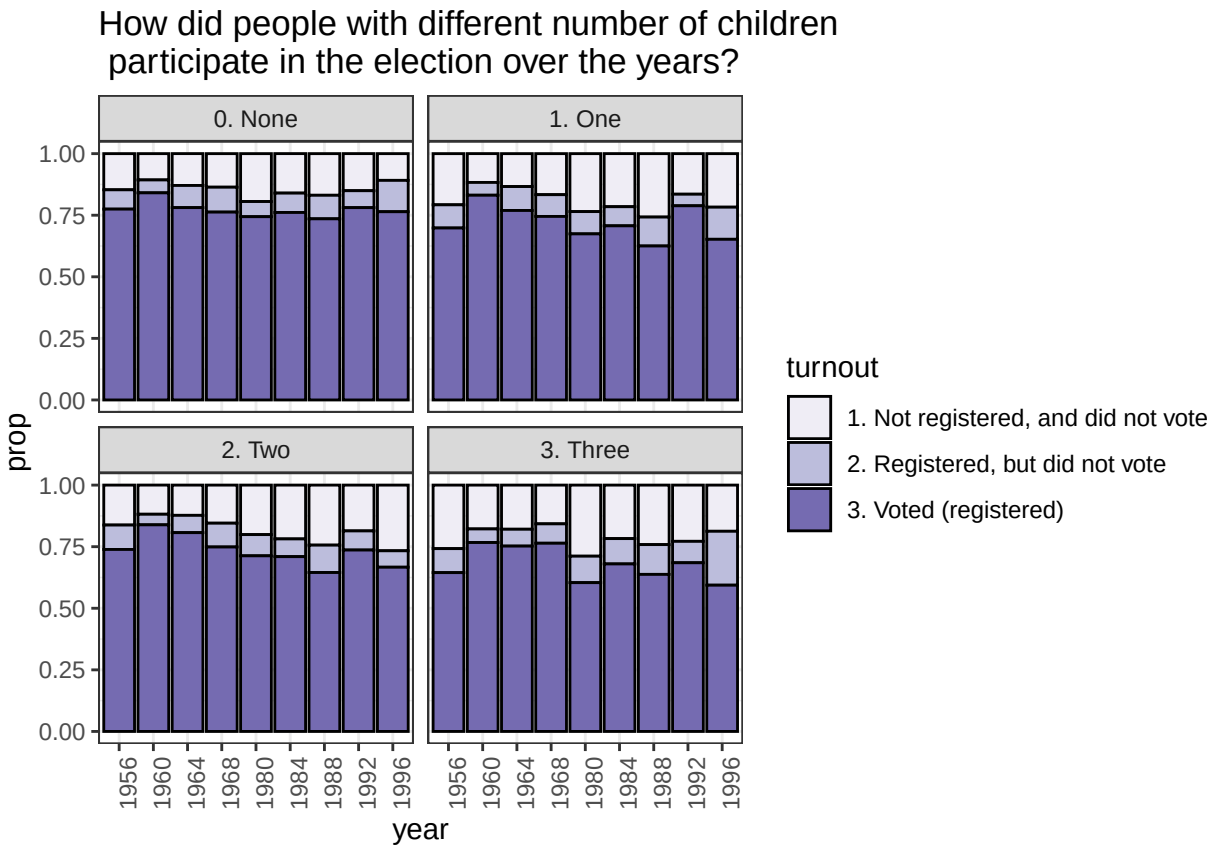
### 5.1 voting due to children they have

```
load(file="../output/data_use.RData")
Election_years=as.character(seq(1952, 2016, 4))

anes_burden=anes_dat%>%
  mutate(
    Number_of_children=as_factor(VCF0138),
    children_under_6=as_factor(VCF0138a),
    children_6_9=as_factor(VCF0138b),
    children_10_13=as_factor(VCF0138c),
    children_14_17=as_factor(VCF0138d),
    social=as_factor(VCF0148a),
    work_st=as_factor(VCF0116),
    work_HH=as_factor(VCF0117),
    vote=as_factor(VCF0706),
    gender=as_factor(VCF0104),
    turnout=as_factor(VCF0703),
    year=as_factor(VCF0004)

  )%>%
  filter(year %in% Election_years)
```

#### 5.1.1 Participating due to number of children they have

We want to use the data to see whether the number of children they have will affect the participation. So we did a bar plot of the participation rate: voted, registered but did not vote, did not register and did not vote. We divided the data into four part due to the number of children. And also, we want to discuss the volatile over the years about these factors.

```
load(file="../output/data_use.RData")
anes_to_children= anes_burden %>%
  filter(!is.na(Number_of_children) & !is.na(turnout))%>%
  group_by(Number_of_children,year)%>%
  count(turnout)%>%
  group_by(Number_of_children,year)%>%
  mutate(
    prop=n/sum(n)
  )

ggplot(anes_to_children,
       aes(x=year, y=prop, fill=turnout)) +
  geom_bar(stat="identity", colour="black")+
  facet_wrap(~Number_of_children, ncol=2) +
  theme_bw()+
  theme(axis.text.x = element_text(angle = 90))+
  scale_fill_brewer(palette="Purples")+
  labs(title="How did people with different number of children \n participate in the election over the
```




From the graph above, we know that people with no child participated most in the election. They have the highest probability of registered and voted. Additionally, they are less volatile over the years. Individuals with three children participated least in the election and they have the highest rate of registered but not voted in 1996. The chart may indicate they care less about the political voting than others which may due to their schedule, gender.

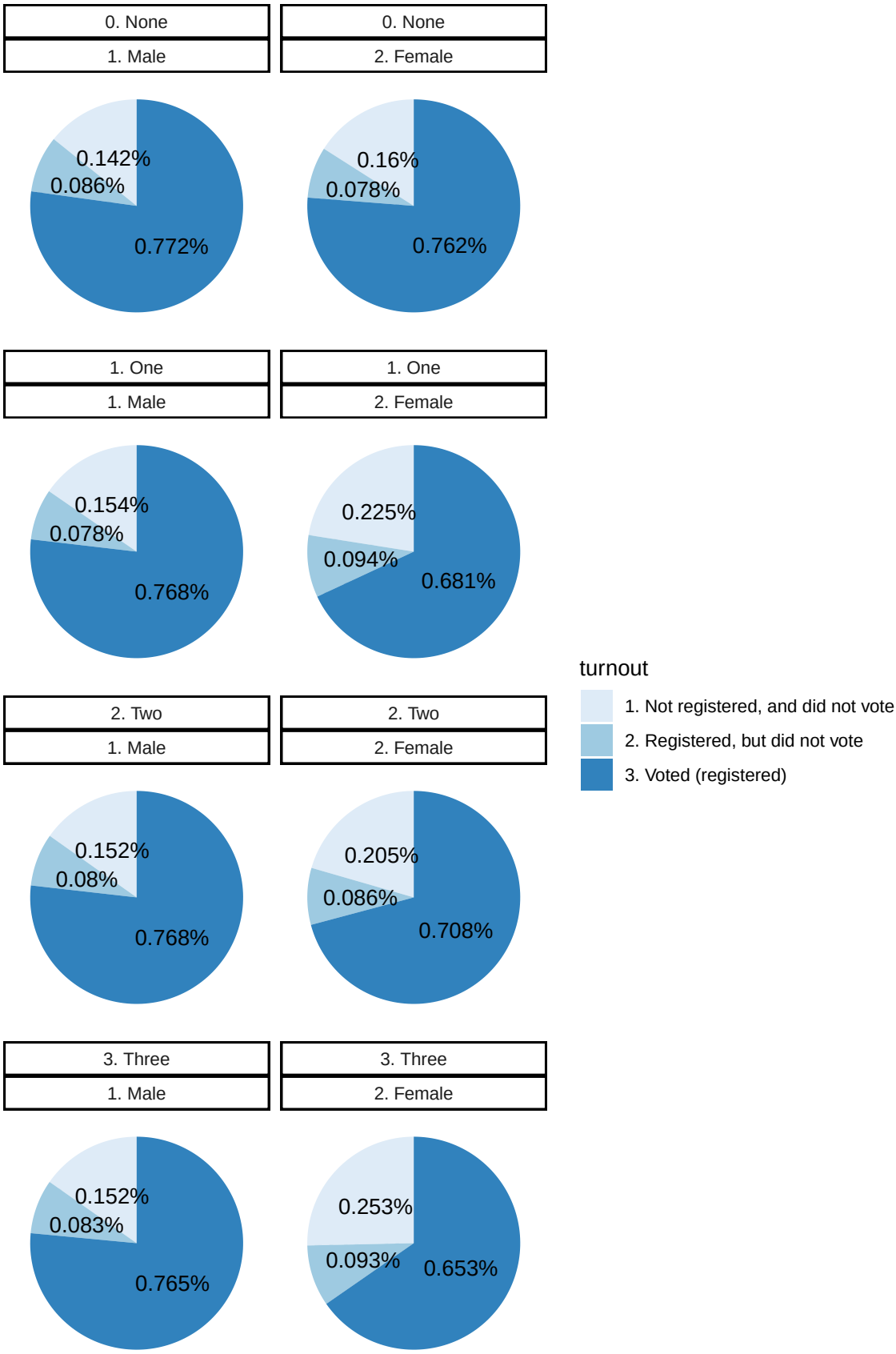
### 5.1.2 Particiation due to gender with same number of children

As we know from previous, the number of children do have a little impact in the participation rate, we are going to see if the gender has impact on individuals with the same number of children. We still grouped the data by number of children but we add the group by gender to see if gender is an impact in participation.

```
load(file="../output/data_use.RData")
anes_to_children1= anes_burden %>%
  filter(!is.na(Number_of_children) & !is.na(turnout) & !is.na(gender) )%>%
  group_by(Number_of_children,gender)%>%
  count(turnout)%>%
  group_by(Number_of_children,gender)%>%
  mutate(
    prop=n/sum(n)
  )

ggplot(anes_to_children1, aes(x="", y=prop, fill=turnout)) +
  geom_bar(stat="identity", width=1) +
  coord_polar("y", start=0) +
  geom_text(aes(label = paste0(round(prop,3), "%")), position = position_stack(vjust=0.5)) +
  labs(x = NULL, y = NULL) +
```

```
  facet_wrap(~Number_of_children+gender, ncol=2) +
  theme_classic() +
  theme(axis.line = element_blank(),
          axis.text = element_blank(),
          axis.ticks = element_blank()) +
  scale_fill_brewer(palette="Blues")+
  labs(title="How did gender and number of children affect on the participation generally?")
```

How did gender and number of children affect on the participation generally?



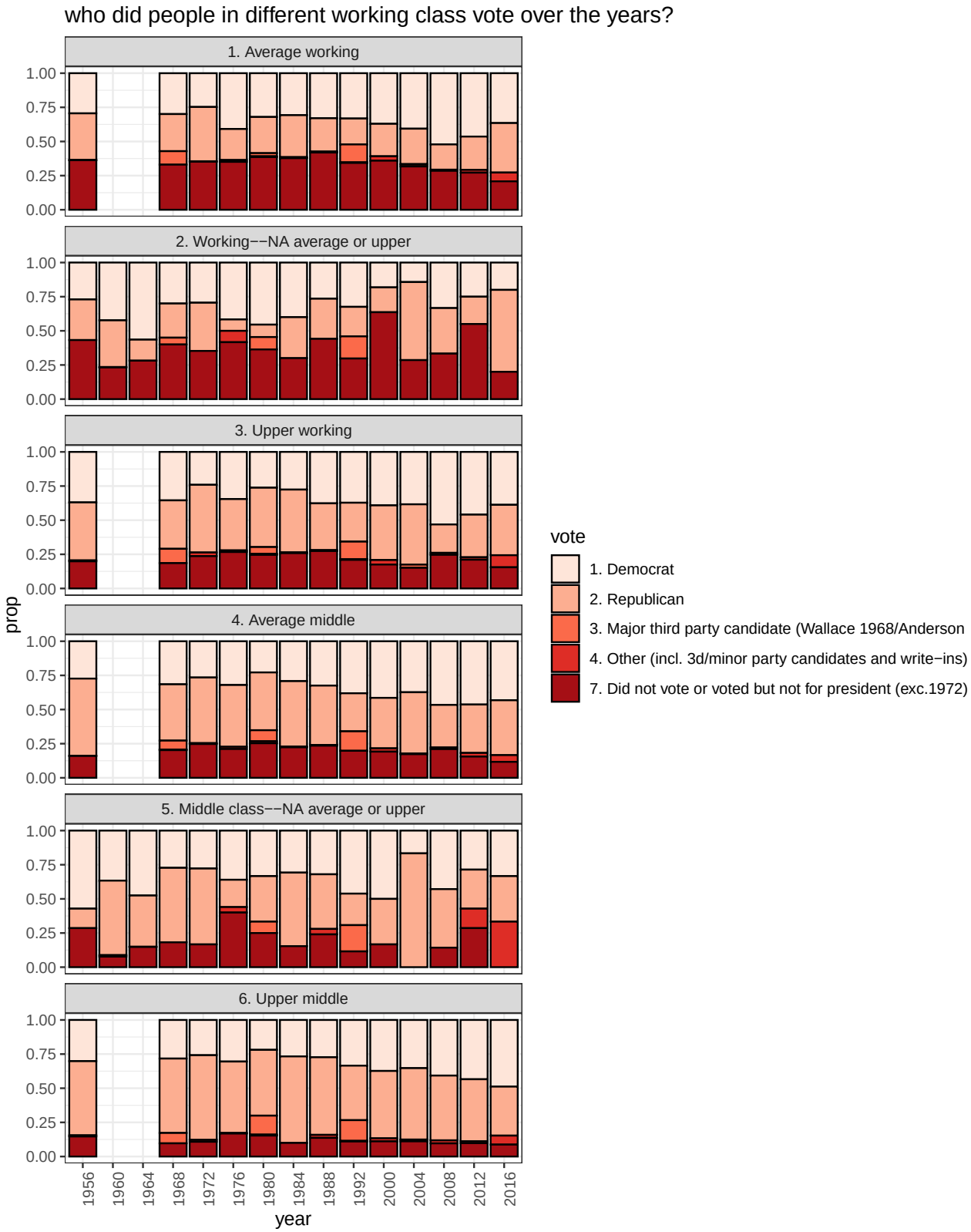

Firstly, we

can conclude from the pie chart that males participate more than females which means males has higher rate of voted than females'. However, there's no big difference between male and female when they have zero child. This may due to they are single and have less burden from family. However, when they have children, especially has one or three children, females have much lower participate rate than males, which means females has lower voted rate, higher rate of not registered and not voted. There are several reasons may cause this situation. Firstly, females are less interested in politics. Secondly, as females have kids, they put more effort in the family and care less about the voting.

## 5.2 voting due to their social classes

The social classes they are in may affect the family burden they are experiencing, which may affect their voting. So we arrange a data set group by social classes. We eliminate all the NA values and we want to figure out how the social class standing will affect the voting.

```
load(file="../output/data_use.RData")
anes_to_social= anes_burden %>%
  filter(!is.na(social) & !is.na(vote))%>%
  group_by(social,year)%>%
  count(vote)%>%
  group_by(social,year)%>%
  mutate(
   prop=n/sum(n)
  )

ggplot(anes_to_social,
       aes(x=year, y=prop, fill=vote)) +
  geom_bar(stat="identity", colour="black")+
  facet_wrap(~social, ncol=1)+
  theme_bw()+
  theme(axis.text.x = element_text(angle = 90))+
  scale_fill_brewer(palette="Reds")+
  labs(title="who did people in different working class vote over the years?")
```




From the distribution graph, we see that for average working class, there are almost more than 1/3 of them didn't vote or voted but not for presidents. And for working class, they have more volatility in voting. They have the most change in voting over the years. And for the upper middle class, they vote more for Republican than others and only a small amount of them didn't vote or voted but not for president.

The individuals in working class and average working class may have bigger family burden than other classes since they have lower salary and they may also need to raise the same amount of children and they may have less welfare and other insurance than the upper class and middle class. So individuals in working class may volatile more than other classes since they want to vote for the most profit one for them. And more individuals in working class didn't vote since they may have no interest in politics when they think voting has no direct profit for them.

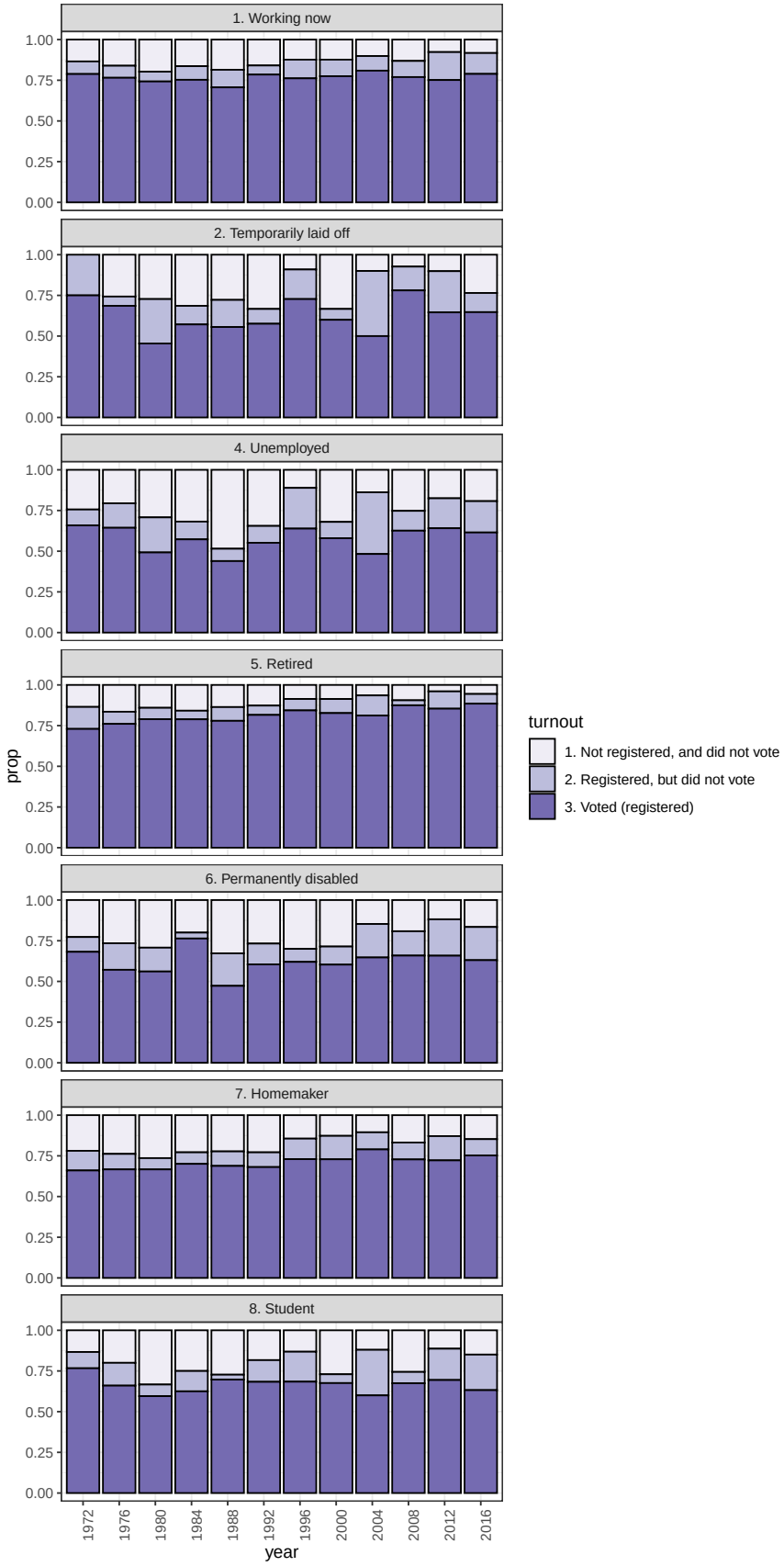
## 5.3 voting due to their work status

### 5.3.1 participation rate due to their working status

Other than social status, working status is also a big part of the family burden. If individuals are unemployed, they may have bigger family burden than employed ones. So we have the following categories, student, homemaker, permanently disabled, retired, unemployed, temporarily laid off, and working now. We want to figure out if the working status has effect on their participation in voting.

```
load(file="../output/data_use.RData")
anes_to_work= anes_burden %>%
  filter(!is.na(work_st) & !is.na(turnout))%>%
  group_by(work_st,year)%>%
  count(turnout)%>%
  group_by(work_st,year)%>%
  mutate(
   prop=n/sum(n)
  )

ggplot(anes_to_work,
       aes(x=year, y=prop, fill=turnout)) +
  geom_bar(stat="identity", colour="black")+
  facet_wrap(~work_st, ncol=1)+
  theme_bw()+
  theme(axis.text.x = element_text(angle = 90))+
  scale_fill_brewer(palette="Purples")+
  labs(title="How did people in different working status participating vote over the years?")
```

How did people in different working status participating vote over the years?




From the chart above, we can see

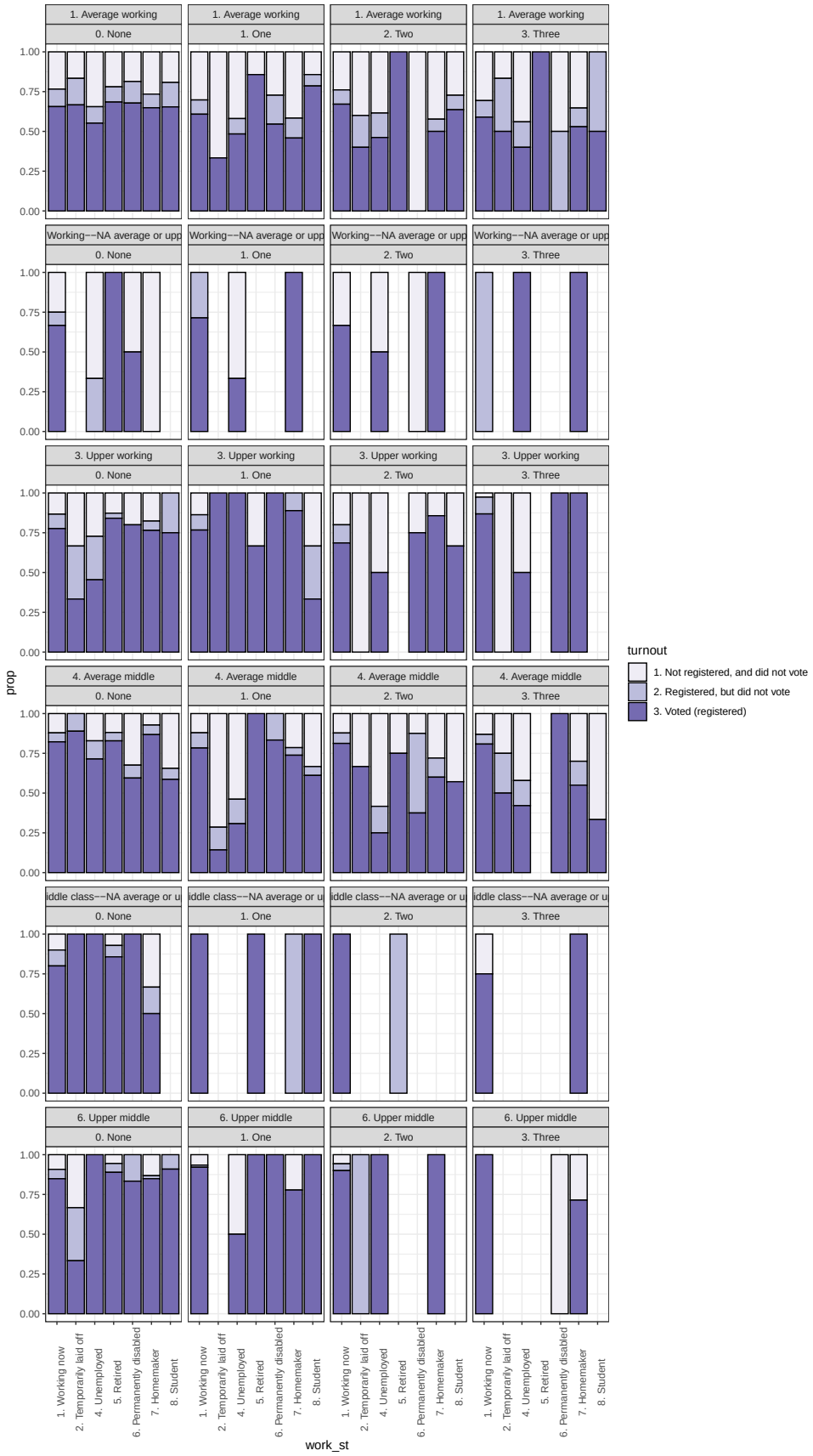
that retired individuals has the highest participation rate in voting, because they may have more spare time and less family burden since they have no children ,they have pension to cover their life, and they don't need to work. Employed and homemakers has the least volatile over the years, because they may have the most stable job and work to do and they do not need to worry too much about the expense in life. Unemployed has the lease participation over the years, since they may need to find a job to solve the burden from family, so they have no time and interest in politics which have no direct effect with their current worries.

## 5.4 Number of children and working status and social status, total effect on participate.

Additionally, we are going to consider all these factors together to get a overall sense of how the family burden affect the participation rate of the voting.

```
load(file="../output/data_use.RData")
anes_to_all= anes_burden %>%
  filter(!is.na(work_st)& !is.na(Number_of_children) & !is.na(social) & !is.na(turnout))%>%
  group_by(work_st,social,Number_of_children)%>%
  count(turnout)%>%
  group_by(work_st,social,Number_of_children)%>%
  mutate(
   prop=n/sum(n)
  )

ggplot(anes_to_all,
       aes(x=work_st, y=prop, fill=turnout)) +
  geom_bar(stat="identity", colour="black")+
  facet_wrap(~social+Number_of_children, ncol=4)+
  theme_bw()+
  theme(axis.text.x = element_text(angle = 90))+
  scale_fill_brewer(palette="Purples")+
  labs(title="How did three possible factos affect participating rate?")
```

How did three possible factos affect participating rate?



For the average working class, the retired people participate the most no matter how many children they have. This may caused since that they have less family burden than others. Their children are all adults and they have pension to cover expenses. So they are more willing to participate in voting.

For the average middle class, they have the most volatile in voting participation due to the working status. For the upper middle class with none child, they have less burden since they are in the upper middle level which means they have no worries about money or hey have at least some savings. The lowest participation in the upper middle class with no children is the temporarily laid off working status, which means they are seeking for a job and have less interest in politics as they are worrying about the job position.

## Other things need to mention

The participation rate between male and female may due to their interest in politics not because of the family burden. Since males are usually more willing to participate in politics than females. The participation rate of the social class may also due to their education level. The participation rate of working status may only reflect the working they have. However, we need to consider some family influence and sponsorship.

## Conclusion

Number of children affect more about females in participation than males and as the increasing of children they have, the less participation rate. Social status is also affect the participation rate, the working class is more unstable than the upper middle class. Working status is also affect the participation rate, retires are more willing to participate than others. Additionally, individuals with higher family burden may participate less in voting.